Enhancing Clustering Technique with Knowledge-based System to Plan the Social Infrastructure Services

Hesham A. Salman[1], Lamiaa Fattouh Ibrahim[2, 3] and Zaki Fayed[4]
[1]*Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
[2]*Department of Computer Sciences and Information, Institute of Statistical Studies and Research, Cairo University, Cairo, Egypt*
[3]*Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia*
[4]*Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, B.P. 42808 Zip Code 21551- Girl Section, Jeddah, Saudi Arabia*

Keywords: DBSCAN Clustering Algorithm, Infrastructure City Planning, Spatial Clustering Algorithm, Urban Planning, Public Service Facility.

Abstract: This article present new algorithm for clustering data in the presence of obstacles. In real world, there exist many physical obstacles such as rivers, lakes, highways and mountains..., and their presence may affect the result of clustering significantly. In this paper, we study the problem of clustering in the presence of obstacles to solve location of public service facilities. Each facility must serve minimum pre-specified level of demand. The objective is to minimize the distance travelled by users to reach the facilities this means also to maximize the accessibility to facilities. To achieve this objective we developed CKB-WSP algorithm (Clustering using Knowledge-Based Systems and Weighted Short Path). This algorithm is Density-based clustering algorithm using Dijkstra algorithm to calculate obstructed short path distance where the clustering distance represents a weighted shortest path. The weights are associated with intersection node and represent the population number. Each type of social facility(schools, fire stations, hospitals, mosque, church…) own many constraints such as surface area and number of people to be served, maximum distance, available location to locate these services. All these constraints is stored in the Knowledge-Based system. Comparisons with other clustering methods are presented showing the advantages of the CKB-WSP algorithm introduced in this paper.

1 INTRODUCTION

Clustering is one of the most useful tasks in data mining process. The different algorithms can be classified regarding different aspects. These methods can be categorized into partitioning methods (Han et al., 2001); (Bradly et al., 1998) hierarchical methods (Zhang and Rousseeuw, 1996); (Guha et al., 1998), density based methods (Ester et al., 1996); (Ankerst et al., 1999), grid based methods (Sheikholeslami et al., 1998); (Agrawal et al., 1998) and model based methods (Kohonen, 1982). The clustering task consists of separating a set of objects into different groups according to some measures of goodness that differ according to the application. The applications of clustering in spatial databases present important characteristics. Spatial databases usually contain very large numbers of points (Nanopoulosl et al., 2001). In spatial databases, objects are characterized by their position in the Euclidean space and, naturally, dissimilarity between two objects is defined by their Euclidean distance (Tan et al., 2006).

Civil engineers often play a major role within the complex planning processes needed to determine the infrastructure location (or layout) and its capacity (Bigotte and Antunes, 2007).

The social infrastructure planning problems faced by public authorities typically consist of determining where the facilities of some infrastructure network should be located and what should be the capacity of these facilities.

Very often, the number of possible solutions for

A. Salman H., Fattouh Ibrahim L. and Fayed Z..
Enhancing Clustering Technique with Knowledge-based System to Plan the Social Infrastructure Services.
DOI: 10.5220/0004391504010408
In *Proceedings of the 5th International Conference on Agents and Artificial Intelligence* (ICAART-2013), pages 401-408
ISBN: 978-989-8565-39-6

social infrastructure planning problems is extremely large and it is advantageous to handle them through a type of optimization model called location (or location-allocation) models. These models are classified as continuous or discrete depending on whether the facilities can be located anywhere on the plane or only in some pre-specified points of the plane. Discrete location models are used more often than continuous location in real-world applications. For this reason they have been extensively studied since the early 1960s, and there is a vast body of literature describing models and solution methodologies.

Clustering technique will be used for helping engineers to determine where the facilities of some infrastructure network should be located and what should be the capacity of these facilities and layout.

In many real applications the use of direct Euclidean distance has its weaknesses (Tan, et al., 2006). The Direct Euclidean distance ignores the presence of streets, paths and obstacles that must be taken into consideration during clustering.

In this paper, a clustering–based solution is presented depending on using the obstructed short path distance and density-Based Clustering techniques.

A typical real-world application of the model is school network planning. In this case, the model would aim to determine the locations and capacities of schools to minimize the distance traveled by students, taking into account that the students must be assigned to the school that is closer to the place where they reside. In section 2 Motivation is discuss. In section 3, the CKB-WSP algorithm is introduced. A case study is presented in section 4. Section 5 discusses related work. The paper conclusion is presented in section 6.

2 MOTIVATION

Spatial clustering algorithms can be classified into four categories. They are the partition based, the hierarchical based, the density based and the grid based. Among all the clustering methods, we found that the partitioning based and density algorithms to be most suitable since our objective is to discover good locations that are hidden in the data. Partitioning based clustering methods include two major categories, k-means and k- medoids. The common premise of these two methods is to randomly partitioning the database into k subsets and refine the cluster centers repeatedly to reduce the cost function. The cost function in the spatial domain is the sum of distance error distance E from all data objects to their assigned centers.

The non center data points are assigned to the centers that they are nearest to it. The k-means algorithm is one of the first clustering algorithms proposed. It is easy to understand and implement, and also known for its quick termination. The k-means algorithm defines the cluster centers to be the gravity center of all the data points in the same cluster. In regular planar space, the cluster gravity center guarantees the minimum sum of distances between the cluster members and itself. However, the research proof (Nanopoulosl et al., 2001) that the characteristic of the gravity center does not behave the same as in obstacle planner space. Instead of representing the clusters by their gravity centers, the k-medoids algorithm chooses an actual object in the cluster as the clusters representative (medoid). Using the real object decreases the k-medoids sensitivity to outliers. This technique also guarantees that the center is accessible by all data objects within the same cluster.

By comparing CLARA and CLARANS with PAM, CLARA first draws random samples of the data set and then does PAM on these samples. Unlike CLARA, CLARANS draws a random sample from all the neighbor nodes of the current node in the searching graph. Efficiency depends on the sample size and a good clustering based on samples will not necessarily represent a good clustering of the whole data. The PAM (Partioning Around Medoids) algorithm, also called the K-medoids algorithm, represents a cluster by a medoid (Tan et al., 2006). Initially, the number of desired clusters is input and a random set of k items is taken to be the set of medoids. Then at each step, all items from the input dataset that are not currently medoids are examined one by one to see if they should be medoids. That is, the algorithm determines whether there is an item that should replace one of the existing medoids. By looking at all pairs of medoids, non-medoids objects, the algorithm choose the pair that improves the overall quality of the clustering the best and exchanges them. Quality here is measured by the sum of all distances from a non-medoid object to the medoid for the cluster it is in.

The total impact to quality by a medoid change TC_{ih} is given by:

$$TC_{ih} = \sum_{h=1}^{k} \sum_{n_i \in C_h} dis(n_h, n_i) \quad (1)$$

An item is assigned to the cluster represented by the medoid to which it is closest (minimum distance or direct Euclidean distance between the customers and the center of the cluster they belong to).

PAM is not suitable for our problem because the

Euclidian distance did not represent the actual in the presence of obstacle. The second reason, the number of facilities is not known before work. The last reason we need each facility service a predefine number of population (density population).

The DBSCAN algorithm is a well-known algorithm of type Density-Based (Yiu and Mamoulis, 2004) algorithm which is used when a cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. This algorithm found clusters with different shapes. This algorithm also is not suitable where our problem has not this property. We search a clustering algorithm which construct cluster with a density within a given range and also a point in this cluster which represent this cluster such that the cost is minimum.

3 CKB-WSP ALGORITHM

The existing of the natural obstacles are affecting on distribution of the service facility on the regions. The responsible civil engineers determine the infrastructure location and layout. Very often, the number of possible solutions for social infrastructure planning problems is extremely large and it is advantageous to handle them through a type of optimization model. These models are classified as continuous or discrete depending on whether the facilities can be located anywhere on the plane or only in some prespecified points of the plane. Discrete location models are used more often than continuous location in real-world applications.

In a certain city, we need to determine the number of public service facility requirements and define their boundaries in such away that satisfy shortest path between users and facilities. We must take into account that each facility must serve a minimum level of demand to be economically viable and that each user must assign to the closest work facility.

The solution we propose in this paper applies to social infrastructure planning problems with the following features:

- The objective is to minimize the demand-weighted total distance (or travel time, or travel cost).
- A facility can only be opened if it serves a minimum level of demand. Thus the capacity of each facility must exceed that given minimum to be economically viable.
- The number of facilities to be opened is an output of the model.
- Users must be assigned to the closest open facility. If the travel distance is the same for two or more different facilities, users should be assigned to one and only one of those facilities.

The problem statement:-

- Inputs:
  - A set T data points $\{t_1, t_2 \ldots t_n\}$ in 2-D map.
  - Surface of area to be plan.
  - Obstacles location.
  - MinPTS = minimum population services for this public services facility.
  - MaxPTS= maximum population that can serviced by this public services facility.
  - Candidate locations of services facilities.
- Objectives:
  - Partitioning the city into k clusters C1, .., Ck that satisfy clustering constraints (minimum and maximum services population) such that the cost function is minimized.

Min TC= $\sum$ obstacle distance (i, j)* wi

Where:

TC is the cost function to be minimize

Obstacle distance (i, j) = min path obstacle distance between node i and facility j of the cluster which is calculate by Dijkstra algorithm

wi = weight of node i = population of node i

Output:

- Optimal number of clusters which satisfy the required objectives.
- locations of public services facility
- boundaries of each cluster.

The proposed algorithm contains three phases. Figure 1 shows the block diagram of the overall system. The following sections describe these three phases.

3.1 Phase I: Pre-planning

The maps used for planning are scanned images obtained by the user from GOOGLE map. It needs some preprocessing operations before it used as digital maps, we draw the streets and intersection nodes on the raster maps, the beginning and ending of each street are transformed into data nodes, defined by their coordinates. The streets themselves are transformed into links between data nodes. The populations are considered to be the weights for each node.

3.2 Phase II: Main-planning Phase

CKB-WSP is divided into two step:

1- Step 1: Preprocessing.
2- Step 2: CKB-WSP algorithm.

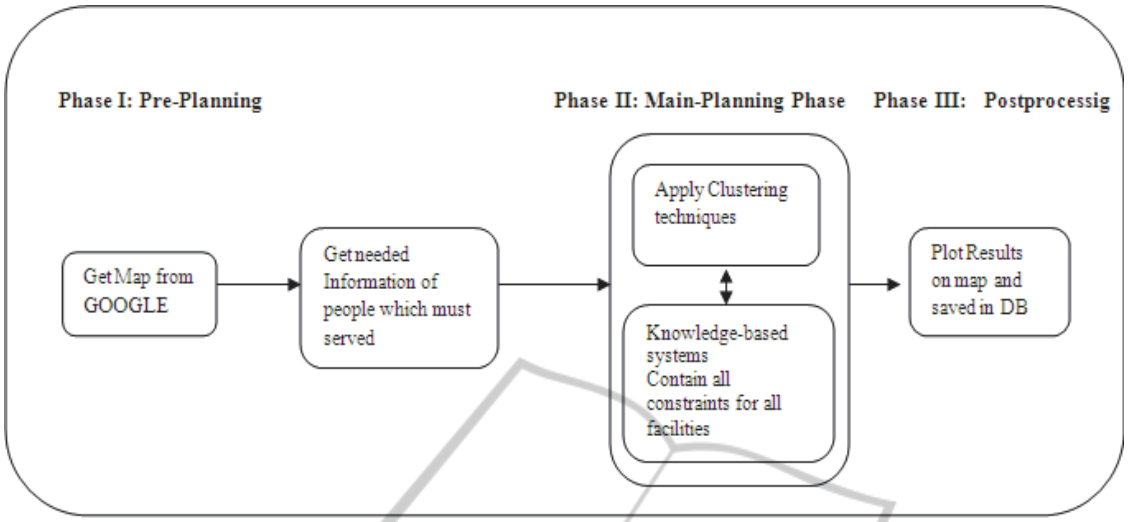

Figure 1: Block diagram of overall the CKB-WSP system.

3.2.1 Preprocessing

During clustering the CKB-WSP often needs to compute the short obstructed path distance between a point and a temporary cluster center. Our aim of pre-processing here is to manipulate information which will facilitate such computation.

CKB-WSP algorithm used Dijkstra algorithm to calculate shortest path from one source to many destinations; in the CKB-WSP we need to calculate the shortest path from public service facility to all nodes (the reason is to determine the suitable location and layout of public service facility) and from one node to all public service facility (the reason is to determine the nearest suitable public service facility that will serve this node). Figure 2 shows the pseudo code of the Dijkstra algorithm**.**

3.2.2 CKB-WSP Algorithm

Figure 3 shows implementation of pseudo code of CKB-WSP algorithm used**.** The algorithm begins with estimate number of clusters which is equal to sum of population of all points divided by MaxPTS, where MaxPTS is the maximum population this facility can served.

The user first inserts the location of candidate. Our package arbitrarily selects K points from candidates to be the initial location of services facilities. After this step the package determines the boundaries of cluster by calculating the obstacle distance from each node to each facility. The algorithm iterate until chooses one which satisfies the conditions and minimize the cost function. Each type of facility has his constraints. These constraints differ from one to another. To open school, government determines two constraints MinPTS which is the minimum number of students that can open a school and MaxPTS the maximum number of students which determine from the surface of the school. The maximum distance which can the students wake to go to school Eps must be know. If the government needs to open mosque, any human to pray need .5m*1 m to know how many humans can pray in this mosque (MaxPTS), divide surface of mosque by .5 m^2. The maximum distance which can the human can wake to go to mosque, Eps, must be know.

```
Function Dijkstra(G, w, s)
For each vertex v in V[G]//
Initializations
      d[v] := infinity
      previous[v] := undefined
      d[s] := 0
      S := empty set
      Q := set of all vertices
While Q is not an empty set // The algorithm
itself
      u := Extract_Min(Q)
S := S union {u}
   for each edge (u,v) outgoing from u
     if d[v] > d[u] + w(u,v)              //
Relax (u,v)
        d[v] := d[u] + w(u,v)
       previous[v] := u
End Function
```

Figure 2: Pseudo code of the Dijkstra algorithm.

Knowledge-Based system is constructed and contain all constraints which is affected the plan of facilities services.

3.3 Phase III: Post Processing

The output mined knowledge is presented graphically and as data in data base. The following section shows case study and demonstrates the output knowledge.

4 CASE STUDY

For real application, the proposed algorithm is applied on a map representing a district in Mecca in Saudi Arabia. This area suffer of mountains the bigger one called Al Nour mountain. The actual map is scanned, then the beginning and the ending of each street are transformed into data points, defined by their coordinates; the streets themselves are transformed into linkages between data points; after that the population of each node is added.

The CKB-WSP algorithm divided the map into convenient number of clusters in which the population is distributed.

Figure 4 shows this area; all dark areas represent mountains area. Figure 5 shows the map after applying CKB-WSP algorithm which divides the map into 7 clusters when Eps= 50 and Minpnt= 4000. Figure 6 shows the map after applying CKB-WSP algorithm which divides the map into 3 clusters when Eps= 70 and Minpnt= 6000. From figure 5 and 6, the location of facilities is not showed in the center of cluster but move towards heavy nodes (population) due to the weight parameter whish is introduced in the cost function.

In the proposed algorithm you can enter in each run the values of the minimum number of population (MinPNT), maximum number of population (MaxPNT) the facility can served and the radius of region (Eps) which the facility will be serve. This makes the algorithm more flexible to plan any facility (e.g. preschool, hospital, telecommunication service……).

5 RELATED WORK

In (Cornuejols et al., 1990), the most common objectives is the minimization of costs. However, in the authors' experience, using this objective in participatory social infrastructure planning processes often poses problems. The main reason is that users are reluctant to accept a cost minimization objective, especially when the matter is the location of facilities as important as schools or hospitals. Another relevant reason is that cost information is often scarce and poor, and cost values can be difficult to estimate (particularly the value of fixed costs).

```
Algorithm CKB-WSP
Input
D={t1, t2, t3,……..tn) / * set of elements
Surface of area to be plan
Obstacles location
Eps maximum distance between I and public
service facility
MinPTS = minimum population services for this
public service facility
MaxPTS= maximum population that can service by
this public services facility
Output
A partition of the D objects into K cluster
Location of public service facility
Boundaries of each cluster
CKB-WSP Algorithm
Enter type of facility
Load constraints parameter from Knowledge-
Based system
Estimate number of cluster= K= (Σ weight of
node I / MaxPTS)
currentTC= big number
Label 1 Arbitrarily select K points from Can-
didate to be the location of services facili-
ties
For (i=1 to candidate No.)
  For (j=1 to number of node)
       Calculate the short path obstacle
  distance from public service   facility (i)
to node(j)
      If (path obstacle  distance < Eps km)
      Then    current population (i)= current
number of population (i) + population of node
 End For
      If (Current population (i) >=     Max-
       PTS)
      Then    add one center of  public ser-
      vice facility and go to label 1
       If (Current population (j) < MinPTS)
        Go to label 1
End for
For each node in the city select the best
    location service for it by calculating the
    obstacle distance between the nodes and
    each location service
Calculate the number of population of each
core
Calculate TC If TC < currentTC than currentTC
= TC go to label 1
Save solution
```

Figure 3: Implementation of CKB-WSP algorithm.

On the contrary, the maximization of accessibility to facilities tends to be a more consensual objective among the different stakeholders. This objective is usually represented in location models by the minimization of the demand-weighted total (or average) distance traveled to obtain the service. One classic model that considers this objective is the *p*-median model (Mirchandani, 1990) originally; the *p*-median model is based upon two important assumptions: one

knows a priori how many facilities should be opened and the capacity of facilities does not have to satisfy maximum and/or minimum limits. However, in a social infrastructure planning problem, the number of facilities to locate is typically one of the desired outcomes (rather than a parameter) and the capacity of facilities must be within certain limits.

Although there is an abundant literature on incapacitated *p*-median models, capacitated versions have been less studied. Moreover, most of the existing models only take into account maximum capacity constraints (recent examples include (Lorena and Senne, 2004); (Ceselli and Righini, 2005); (Diaz and Fernandez, 2006). However, minimum capacity constraints are important because they model the minimum level of demand that facilities must satisfy to be economically viable. Above this level, possible economies of scale have already been made, and unit facility costs can be considered to be constant.

Figure 4: Area in Macca City in Saudi Arabia.

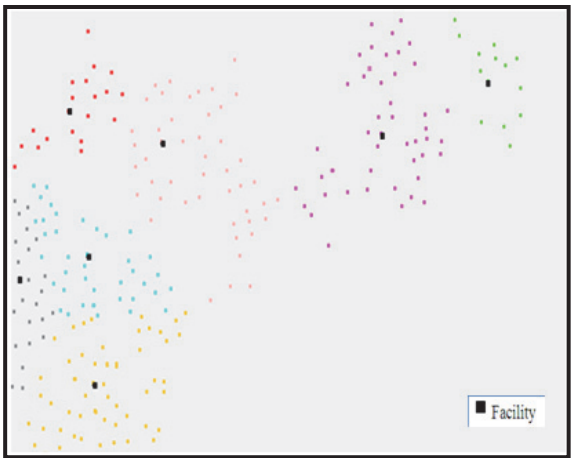

Figure 5: Using CKB-WSP algorithm considering the location of obstruct Eps= 50 and Minpnt= 4000.

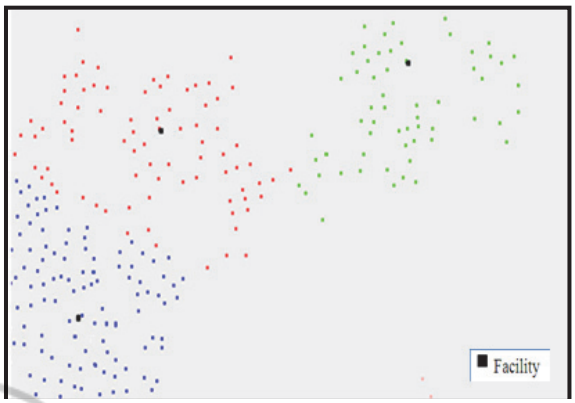

Figure 6: Using CKB-WSP algorithm considering the location of obstruct Eps= 70 and Minpnt= 6000.

In (Bigotte, 2007), the model solve small-size instances exactly using exact method. For large-size instances, it uses heuristic methods. First, Using a classic local search heuristic (Add + Interchange) and classic population heuristic (GA) failed to identify optimum or near-optimum solutions but the solutions provided by the (Add + Interchange) heuristic were better than those given by the GA. Second, Using, Tuba Search TS and Specialized Local Search Heuristic SLSH in which the neighborhood structure was improved to better represent the specific features of the model. SLSH generally provides better solutions than TS, though requiring a larger computing effort.

Table 1 described the different comparison between the proposed method and other methods using in facilities planning. There are two methods that are frequently used here: Tabu search (Bigotte and Antunes, 2007)and Genetic Algorithms (Bigotte and Antunes, 2007).

6 CONCLUSIONS

Clustering analysis is one of the major tasks in various research areas. The clustering aims, to identify and extracting significant groups in underlying data. Based on certain clustering criteria; the data are grouped so that the data points in a cluster are more similar to each other than points in different clusters. In this paper, we introduced a clustering solution to the problem of locate public Service facility in the presence of physical obstacles, the CKB-WSP algorithm. This algorithm is density-based clustering algorithm using distances which are weighted shortest obstacle path distance (not Euclidian distance) and satisfying facilities constraints due to use of knowledge-based system. The result is a realistic

solution representing the population demand with minimum costs due to modify in cost function.

The application of the CKB-WSP algorithm was illustrated through a case study in a location of districts in Mecca in Saudi Arabia. Experimental results and analysis indicate that the CKB-WSP algorithm is effective to satisfy populations demand for facility constructed in an area where population is non-homogeneous due to the presence of obstacles.

The existence of Knowledge-Based System helps us to plan any new facility serves after define the constraints of this facility in the Knowledge-based.

REFERENCES

Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan,P., (1998). Automatic subspace clustering of high dimensional data for data mining application. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*.

Ankerst, M., Breunig, M., kriegel, H. and Sander, J., (1999). OPTICS: Ordering points to identify the clustering structure. In *Proc. 1999 ACM-SIGMOD Int. Conf. Management of data (SIGMOD'96)*.

Bigotte, J. F., Antunes, A. P., Social Infrastructure Planning: A Location Model and Solution Methods, *Computer-Aided Civil and Infrastructure Engineering* 22 (2007) 570–583.

Bradly, P., Fayyad, U., and Reina, C., (1998). Scaling clustering algorithms to large databases. In *proc. 1998 Int. Conf. Knoweldge Discovery and Data mining*.

Ceselli, A.,&Righini,G. (2005),A branch-and-price algorithm for the capacitated p-median problem, *Networks*, 45(3), 125–42.

Cornuejols, G., Nemhauser, G. L. &Wolsey, L. A. (1990), The uncapacitated facility location problem, in P.B. Mirchandani and R. L. Francis (eds.), *Discrete Location Theory*, JohnWiley & Sons, New York, pp. 119–71.

Diaz, J. A. & Fernandez, E. (2006), Hybrid scatter search and path relinking for the capacitated p-median problem, *European Journal of Operational Research,* 169(2), 570– 85.

Ester, M., Kriegel, H., Sander, J. and Xu, X., (1996). A density based algorithm for discovering clusters in large spatial databases. In *Proc. 1996 Inc. Conf. Knowledge discovery and Data mining (KDD'96)*.

Guha, S., Rastogi, R., and Shim, K. (1998). Cure : An efficient clustering algorithm for large databases. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*.

Han, J., and Kamber, M., *Data Mining Concepts and Techniques*, Elsevier, 2011.

Han, J., Kamber, M., and Tung, A., (2001). *Spatial Clustering Methods in data mining: A Survey, Geographic Data Mining and Knowledge Discovery*.

Ibrahim, L. F., (2011) Enhancing Clustering Network Planning Algorithm in the Presence of Obstacles, *KDIR International Conference on Knowledge Discovery and Information Retrieval, KDIR is part of IC3K, the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Paris, France 26- 29 October 2011.

Kohonen, T.,(1982). Self organized formation of topologically correct feature map. *Biological Cybernetics*.

Lorena, L. A.N.&Senne, E. L. F. (2004), A column generation approach to capacitated p-median problems, *Computers and Operations Research*, 31(6), 863–76.

Mirchandani, P. B. (1990), The p-median problem and generalizations, in P. B. Mirchandani and R. L. Francis, (eds.), *Discrete Location Theory*, John Wiley & Sons, New York, pp. 55–117.

Nanopoulosl, A., Theodoridis , Y., Manolopoulos, Y., (2001). C2P: Clustering based on Closest Pairs. *Proceedings of the 27th International Conference on Very Large Data Bases*, p.331-340, September 11-14.

Sheikholeslami, G., Chatterjee, S. and Zhang, A., (1998). Wave Cluster : A multi- resolution clustering approach for very large spatial databases. In *Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97)*.

Tan, P., Steinback, M., and Kumar, V., (2006). *Introduction to Data Mining*. Addison Wesley.

Yiu, M., Mamoulis, N. (2004). Clustering Objects on a Spatial Network. *SIGMOD Conference* p 443-454.

Zhang, T., Ramakrishnan, R., and Livny, M., (1996). *BIRCH: an efficient data clustering method for very large*.

Table 1: Relative works.

Algorithm	Input Parameters	Results	Location of facility	Constraints	Type of distance	Consider obstacles
Genetic	- Data points - Population - Initial probability - Mutation probability - Crossover probability - Number of iteration - Selection pressure	# of facilities	Optimal placement	Fitness Function	Euclidean distance	NO
Tabu Search	- Data points - Init probability - Generation probability - Recency factor - Frequency factor - Number of iteration - Number of neighbors	# of facilities	Optimal placement	Tabu List	Euclidean distance	NO
CKB-WSP	- Data points -Eps - MinPNT - MaxPNT	- Core points -# of facilities	Core point	MinPts, Eps and obstacles	Short path obstacles distance	YES